UNIVERSITY OF COPENHAGEN

De Novo Discovery of Structured ncRNA Motifs in Genomic Sequences

Ruzzo, Walter L; Gorodkin, Jan

Published in:
RNA Sequence, Structure, and Function

DOI:
10.1007/978-1-62703-709-9_15

Publication date:
2014

Document version
Early version, also known as pre-print

Citation for published version (APA):
Ruzzo, W. L., & Gorodkin, J. (2014). *De Novo* Discovery of Structured ncRNA Motifs in Genomic Sequences. In J. Gorodkin, & W. L. Ruzzo (Eds.), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods* (Vol. 1097, pp. 303-18). Methods in molecular biology (Clifton, N.J.) https://doi.org/10.1007/978-1-62703-709-9_15

Download date: 04. May. 2023

Chapter 15

De Novo Discovery of Structured ncRNA Motifs in Genomic Sequences

Walter L. Ruzzo and Jan Gorodkin

Abstract

De novo discovery of "motifs" capturing the commonalities among related noncoding structured RNAs is among the most difficult problems in computational biology. This chapter outlines the challenges presented by this problem, together with some approaches towards solving them, with an emphasis on an approach based on the CMfinder program as a case study. Applications to genomic screens for novel *de novo* structured ncRNAs, including structured RNA elements in untranslated portions of protein-coding genes, are presented.

Key words CMfinder, Mutual information, ncRNA discovery, ncRNA gene, ncRNA motif, Riboswitch

1 Introduction

De novo discovery of "motifs" capturing the commonalities among related noncoding structured RNAs is among the most difficult problems in computational biology. The fundamental problem is that structurally constrained RNAs evolve while conserving structure, not sequence. Thus, computationally expensive structure prediction and/or structure-based search algorithms somehow must be part of the equation, to winnow rare homologous ncRNAs from the chaff of large genomes. Successful approaches to date combine sensitive algorithms for motif discovery with homology search and exploit prior biological knowledge wherever possible.

This chapter outlines this problem. Our goal is threefold—to illustrate the challenges presented by this problem, to point out some approaches that have been partially successful in addressing them, and to highlight areas where further improvements are especially desirable. As a particular case study, we emphasize an approach based on the CMfinder program [1], successfully used for discovery of functional noncoding RNA elements in prokaryotes (e.g., [2, 3]) and of strong candidates in vertebrates (e.g., [4]).

Jan Gorodkin and Walter L. Ruzzo (eds.), *RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods*, Methods in Molecular Biology, vol. 1097, DOI 10.1007/978-1-62703-709-9_15,
© Springer Science+Business Media New York 2014

Our outline is certainly not comprehensive. Other methods have been applied for genome-wide screens using different alignment strategies (ranging from making use of sequence-based alignments to complete realignment by Sankoff-style approaches), different scoring schemes, different exploitation of biological knowledge, etc. For more, we refer interested readers to other chapters in this volume and to any of the many excellent review articles that have appeared recently, e.g., [5–13].

2 Problem Overview

The comparative method, also known as covariation analysis, is one of the most powerful tools available for the elucidation of RNA secondary structures; *see*, e.g., Chapter 16, or [14, 15]. The key point is that mutations destroying a structurally important RNA base pair by altering one of its nucleotides may be repaired by a *compensating* change to its partner; observations of such changes between homologs thus highlight the base-paired regions in these molecules. The technical challenges in exploiting this are to find and align (sufficiently many) homologs and to extract the structural signal from them. Having sequences at appropriate evolutionary distances is critical to this, since very similar sequences, although easily recognizable and alignable, will exhibit few compensating substitutions; conversely, highly diverged sequences may exhibit many examples of compensatory changes, but may be difficult to find and difficult to align (either in sequence or structure).

In short, for genome-scale discovery of structured RNAs, success hinges on (a) selecting the right input data, (b) aligning it well (with attention to potential RNA secondary structure), and (c) inferring the important shared features of the alignment, including structure. Implicit in this is that (d) *local* alignment is critical, since exact boundaries of structured elements are unlikely to be known a priori, (e) entire input sequences may need to be discarded, for essentially the same reason, and (f) an iterative scheme involving genome-scale homology search for additional elements matching the discovered motif is valuable, since the initial input is unlikely to be complete, and additional examples will allow construction of more accurate motif models, iteratively amplifying discovery.

As a case study, this chapter focuses on one approach to the automation of this process, based on the CMfinder algorithm [1]. We do not suggest that this approach is the last word on the problem—it is not. But it is one of several available successful starting points. Many other tools may substitute for particular steps in the process, and we hope improvements will be made to all of them, and that the discussion below will help illustrate what may work and what needs improving.

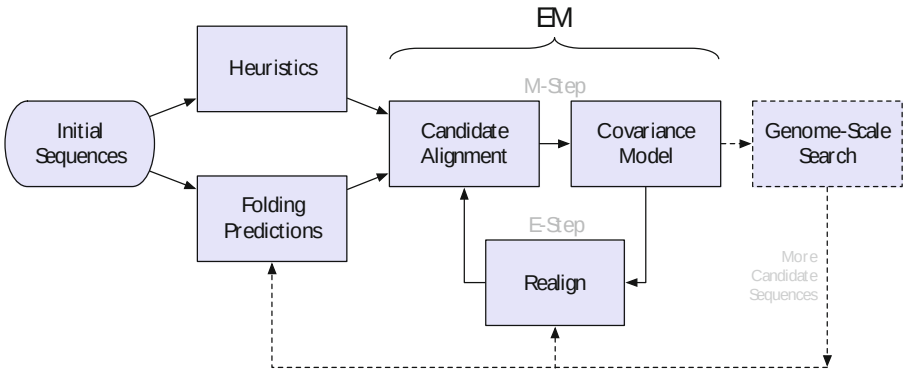

Fig. 1 Overview of the CMfinder discovery pipeline

As an example application, consider the problem of riboswitch discovery. *Riboswitches*, first discovered in 2002, are domains typically found in 5′ untranslated regions of messenger RNAs, where they conditionally control gene expression based on the presence of specific small molecules [16]. Ideally, one would like to have a tool that, given one or more genomes, would identify all riboswitches in them. Unfortunately, this is well beyond the state of the art; functional RNAs are simply too diverse and/or rare to be mechanically identified *de novo* amidst the vast bulk of non-RNA-containing genome sequence, at least by current methods. However, we can let discovery be guided by the known biology. Given that riboswitches are typically *cis*-acting elements with well-conserved secondary structures that regulate specific biochemical pathways in phylogenetically related prokaryotes, one might hope to find examples by examining upstream sequences extracted from orthologous genes in specific bacterial clades. Furthermore, given that multiple copies of a specific type of riboswitch sometimes regulate multiple steps in a particular biochemical pathway (steps *not* performed by orthologous enzymes), search based on the motif identified upstream of some enzyme in a pathway may well reveal additional paralogous riboswitch instances upstream of other enzymes in the same pathway, thus refining the motif model and providing further support for the significance and function of the RNA.

Figure 1 outlines the process. Input sequences that might contain a functional RNA motif, say 1,000 base pair sequences upstream of the start codons of orthologous enzymes in some prokaryotic clade, are gathered. A collection of smart heuristics applied to these input sequences, based on both sequence conservation and single-sequence structure prediction (Chapter 4, [17, 18]), results in a candidate alignment, from which a consensus RNA secondary structure prediction and covariance model (Chapters 5, 8 and 9, [19]) are built. This initial model is refined by an Expectation-Maximization-like (EM-like) [20, 21] iteration in

which the input sequences are aligned to the model, then the model is rebuilt from the refined alignment. The resulting covariance model may be used for genome-scale search [19], hopefully uncovering additional instances of the ncRNA motif, as suggested in the riboswitch example, which may then be integrated into the model building process, hopefully further improving sensitivity and specificity.

The remainder of this chapter will describe these algorithmic tools in more detail and summarize some of the results obtained using them.

3 The CMfinder Algorithm

CMfinder builds on the DNA motif finding program MEME [22] in using an EM framework to search for motif instances embedded in a simple background, but replaces MEME's ungapped position weight matrix motif models with Covariance Models (CMs, [19, 23, 24], and Chapters 5, 8 and 9) to describe RNA motifs. As described above, it contains three main components: initial heuristic alignment, covariance model inference (the "M-step"), and CM-based realignment (the "E-step"). We will describe each in turn. (As usual, the synopsis below omits certain important details, but hopefully captures some of the more interesting features of the methods.)

3.1 Heuristic Alignment

This step identifies the approximate location and structure of a motif. A key issue is the tradeoff between accuracy and efficiency, but noting that the motif will be refined later, alignment errors are tolerable, provided that good alignments are well represented.

To start, strong candidates, i.e., segments with potentially stable secondary structure, are identified by using a single-sequence folding program [18] to compute the minimum free energy of all subsequences of the input. Candidates are ranked by their free energy, scaled by sequence length.

Local regions of sequence conservation are found by BLAST search, and pairs of candidates from different sequences are selected for comparison, using the structural alignment heuristic sketched below, if they are compatible with these "BLAST anchors." This heuristic improves accuracy by preventing obvious misalignments, as well as saving time by reducing the number of structural alignments calculated. The overall initial alignment consists of a central "consensus" candidate along with its nearest match in each other sequence. This process is repeated on the unselected candidates to find several initial alignments as seeds for the EM iteration.

Candidates are compared using the tree-edit algorithm of Hofacker et al. [18] (also *see* Chapter 12), modified to compare

both unpaired and paired nucleotides. This improves discrimination among RNAs with relatively simple structures. This is a simpler comparison heuristic than those used in Carnac [25], ComRNA [26], or Sankoff-style algorithms (Chapter 13 or [27–29]), for example. Its drawbacks include potentially inaccurate secondary structure prediction and the simplified edit distance model, but it is relatively fast (approximately quadratic in the length of the candidates).

3.2 Model Inference

The key component of the M-step of the EM iteration is to (re-)build the covariance model to maximize the likelihood of the input data given the model. Much of this process is similar to analogous procedures from COVE [23] and CMbuild from Infernal [19]: given the estimated motif positions in the current alignment, and Dirichlet priors, maximum likelihood estimates for the transition and emission probabilities for the CM state machine are obtained, basically by estimating transitions/emissions observed on the training data. One technical innovation that we outline below is a Bayesian formulation of consensus secondary structure prediction that smoothly blends thermodynamic structure prediction with mutual information (defined in, e.g., Chapter 1, Eq. 3). The former is especially useful when sequence identity is high/covariation is low and the later is valuable when the opposite is true.

For $1 \leq i \leq l$, let L_i be the ith column of the current (length l) alignment $D = (L_1, L_2, \ldots, L_l)$ of n sequences, and let $\sigma = (\alpha, \beta)$ be the consensus secondary structure for D, where α is the set of indices of unpaired columns and β is the set of pairs of indices of base paired columns. The goal is to find the structure $\hat{\sigma}$ that maximizes $P(D, \sigma)$, the joint likelihood of the alignment and structure. Assuming independence of columns and column pairs,

$$P(D|\sigma) = \prod_{k \in \alpha} P(L_k) \prod_{(i,j) \in \beta} P(L_i L_j)$$

$$= \prod_{1 \leq k \leq l} P(L_k) \prod_{(i,j) \in \beta} \frac{P(L_i L_j)}{P(L_i) P(L_j)}.$$

The likelihood of an observed column is $P(L_i) = \prod_{x \in \{A,C,G,U\}} (p_x)^{n_x}$, where p_x is the probability of observing x in a row of L_i, and n_x is the observed number of such rows (ignoring gaps and assuming sequences are independent, i.e., a deep "star" phylogeny). Noting that n_x/n is the maximum likelihood parameter estimate for p_x (based on a multinomial model of the observed data L_i), and using an analogous expression for $P(L_i L_j)$, one can show that the term $I_{ij} = \log \frac{P(L_i L_j)}{P(L_i) P(L_j)}$ is proportional to the mutual

information between columns i and j. The optimal structure $\hat{\sigma}$ maximizes $\sum_{(i,j)\in\beta} I_{ij}$. This approach is used by COVE [23] and works well in large, phylogenetically diverse datasets, but less well when covariance is limited, as with a few closely related sequences.

CMfinder introduces an informative prior on structures. If s_i is the prior probability that column i is single stranded, and p_{ij} the prior that columns i and j are base paired, then $P(\sigma) = \prod_{k\in\alpha} s_k \prod_{(i,j)\in\beta} p_{ij}$, and $P(D,\sigma)$ becomes:

$$P(D,\sigma) = P(D|\sigma)P(\sigma) = \prod_{1\le k\le l} P(L_k)s_k \prod_{(i,j)\in\beta} \frac{P(L_iL_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{s_is_j}.$$

The maximum likelihood structure σ maximizes $K = \sum_{(i,j)\in\beta} K_{ij}$ where

$$K_{ij} = \log \frac{P(L_iL_j)}{P(L_i)P(L_j)} \frac{p_{ij}}{s_is_j} = I_{ij} + \log \frac{p_{ij}}{s_is_j},$$

and a simple dynamic programming algorithm can choose a compatible, pseudoknot-free set of base pairs maximizing K.

CMfinder's prior on structures is based on a thermodynamic model. For each sequence, calculate the partition function P_{ij} (Chapter 4, [18, 30]), which estimates the probability of forming base pair i,j, averaged over all possible structures. The column pairing probabilities p_{ij} are estimated by averaging the partition functions of the aligned sequences, and s_i is estimated as $1 - \sum_j p_{ij}$.

Since p_{ij} and s_i are data-dependent, they are not "priors" in a strict Bayesian sense. However, the mutual information and the partition function look at the same data in different ways: mutual information measures the conservation of covarying base pairs in the particular sequences from an evolutionary perspective, while the partition function uses a thermodynamic model that is generically applicable to all RNAs. Combining them leverages both approaches: the energy model dominates when there is little mutual information and conversely mutual information dominates when the thermodynamic predictions are ambiguous. RNAalifold [31] uses a similar approach, calculating a linear combination of free energy and mutual information. (Since the partition function is proportional to $\exp(\Delta E/kT)$, free energy and log probabilities are comparable quantities.)

Conceptually, any of the many other RNA alignment/folding programs might be substituted for the particular method outlined here; *see*, e.g., Chapters 7, 13 and 14 and the references therein.

3.3 Realignment

The purpose of the E-step in the EM framework is to find the expected values of hidden parameters, which, in this problem, define the position (if any) of the motif instance in each input sequence as well as its alignment to the motif consensus. These values implicitly weight the candidates considered in the M-step, so that model rebuilding emphasizes good matches over poor ones. CMfinder accomplishes this by "scanning" each input sequence using the covariance model to identify the highest scoring subsequences of each input sequence, and aligning each to the model via the highest-probability path through the model (variously called the Viterbi or CYK algorithm; *see* Chapters 5, 8 and 9).

As noted earlier, an important aspect of the overall pipeline is the ability to use the model to search for additional matches in new genome sequences, and to incorporate them into the model. This fits naturally in the scheme outlined above—any high-scoring matches found in a genome scan can be aligned to the model by exactly the same process given in the previous paragraph, and (after calculating their partition functions [18, 30]) the model rebuilt as described in Subheading 3.2. This feature has proven highly effective at discovering large and diverse RNA families from a small set of related sequences.

3.4 Motif Scoring

One further piece of the puzzle is scoring—how does a novel "motif" stack up against ones discovered from *random* genomic sequences? A variety of approaches have been proposed to answer this question. One important class of solutions is *shuffling*—whatever scoring approach is taken, score the "real" motif by that method and compare it to similarly generated scores from a large number of sequences formed by randomly shuffling the nucleotide sequence of "real" motifs. An important *caveat* here is that, since the stacking energy in RNA helices is nucleotide-dependent, the average *dinucleotide* composition matters, so appropriate shuffling procedures must preserve these quantities. Altschul and Erickson [32] demonstrate how to shuffle single sequences while exactly preserving dinucleotide statistics, and methods that approximately achieve this goal for multiple sequence alignments are available [33–35].

Alternatively, a phylogenetically informed approach is possible—given a phylogenetic tree capturing the observed motif instances, together with branch lengths and estimated rates of indels, nucleotide substitutions (in unpaired regions) and nucleotide-pair substitutions (in paired regions), are the changes observed in the inferred motif more likely according to a model that accounts for base pairing or one that treats all columns as independent? A number of programs incorporate such models for phylogenetic inference (e.g., RAxML [36], phase [37]) and

others directly use them for motif scoring (e.g., pfold [38, 39], EvoFold [40], pscore [41]). *See* also Chapter 16.

A large number of other approaches have been explored, including the scoring scheme used in RNAalifold [31, 42], and the SVM regression scheme used in RNAz [43].

4 Applications to De Novo Discovery of ncRNA Motifs

Our main application of interest is discovery of structured RNA motifs. These can be part of novel ncRNA genes or, for example, regulatory structures in untranslated regions of mRNAs. As with all methods, they are benchmarked on existing data, which, sadly, are sparse. It is therefore important to recognize that evaluating and comparing methods of this sort is fraught with difficulties, and conclusions depend strongly on specific benchmark data and/or evaluation criteria. Nevertheless, it seems safe to say that the CMfinder pipeline has been successful in several contexts.

The original CMfinder paper [1] demonstrated good results on a variety of Rfam [44–47] families, in terms of accurate recovery of the accepted consensus structure, robustness to inclusion of extraneous flanking sequences and to a declining proportion of motif-bearing sequences in its input. Yao et al. [2] examined extension of the method to incorporate results of genome-scale homolog search as outlined above, and applied it on a broad scale in 44 Firmicute bacteria. The motivating problem was riboswitch discovery. As suggested in Subheading 2, the starting points for motif discovery were datasets containing a few hundred nucleotides of unaligned noncoding sequence upstream of the start codons of genes containing orthologous protein domains, as identified by NCBI's Conserved Domain Database (CDD) [48]. Of the 13 *cis*-regulatory Rfam families present in the clade, 11 (mostly riboswitches) appeared among the 50 top-ranked motifs produced by this automated process, and the resulting models achieved greater than 75% sensitivity and specificity both in identifying family members, and in identifying paired nucleotides in them. The results also showed good rejection of negative controls (permuted alignments) and good recovery of RNA motifs known in the literature but not then in Rfam. Additionally, it discovered a number of novel elements such as ribosomal protein leaders that were consequently added to Rfam. Weinberg et al. [3] carried this analysis into other phyla, characterizing 22 strong candidate *cis*-regulatory RNAs, of which at least 5 were subsequently verified to be riboswitches [49–53]. Extensions of these techniques also have been used for the discovery of other riboswitch and ncRNA candidates; *see*, e.g., [54, 55].

In another direction, Torarinsson et al. [4] applied CMfinder to the human genome. Specifically, they ran it on MULTIZ alignment blocks [56] provided by the UCSC browser [57] within the pilot ENCODE regions of the 17-way human genome alignments [58]. In this context, the alignments were used only to indicate orthology—the detailed nucleotide-level alignments were ignored by CMfinder. The scan identified several thousand candidates, albeit with a high estimated false discovery rate (again based on shuffled alignments). The candidates showed highly significant enrichment for co-occurrence with "indel purified segments" [59]—noncoding segments that appear to be under purifying selection to exclude indels. Functional characterization of the candidates was not attempted, but of a small number of candidates selected for experimental follow-up, most were clearly expressed, usually in a tissue-specific manner.

One very interesting aspect of the results relates to alignment. Whole-genome alignments, as expected, strive to optimize nucleotide identity among aligned positions of different genomes. As explained earlier, however, *covariation* between aligned positions is a crucial indicator of RNA structure. Thus, it is expected that evidence for shared RNA structures is blunted in sequence-based alignments, but the impact of this bias is unclear. As one benchmark, Gardner et al. [60] report that all tested alignment methods exhibit a dramatic degradation in the quality of sequence-based multiple alignments of structured RNAs when sequence identity falls below ≈ 60%. Given that the majority of input alignments considered in [4] fall below this threshold, the effect of these misalignments on genome-scale RNA structure prediction may be quite significant. To further examine this issue, the data from [4] reproduced in Fig. 2 plots the percentage realignment within CMfinder candidates versus the percent sequence identity in the input alignment. It clearly shows the expected trend that candidates found in alignments with lower average identity tend to be more extensively realigned by CMfinder than those from high-identity blocks. More interesting is the extent of the effect. Approximately one-quarter of the candidates were realigned more than 50%, and significant adjustment occurs even in high-identity alignments, where one might have expected that most compensatory changes would be bracketed by well-conserved patches, constraining the changes to be correctly aligned.

To further emphasize the impact of alignment, Torarinsson et al. [61] report thousands of candidate RNA structures shared between human and mouse in regions that whole-genome alignment tools refuse to align. In short, conservation of secondary structure is potentially seriously underestimated by studies based on sequence-only multiple alignments.

As noted earlier, the above approach is only one of many possibilities that have been tried. Both sequence-based and

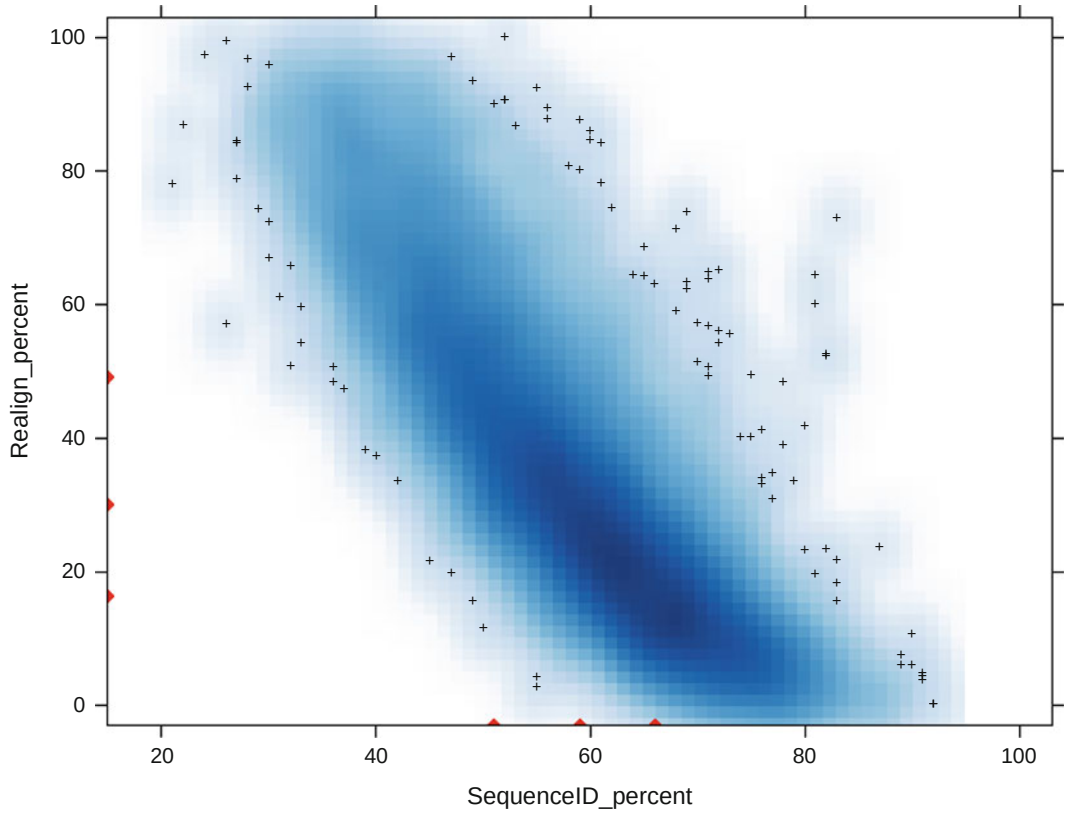

Fig. 2 Realignment versus sequence identity, based on ncRNA candidates from [4]. Shading represents a smoothed density estimate for the scatter plot; pluses mark the 1% of the data points comprising the lowest density regions. Triangles on the axes mark quartiles in the corresponding data

structure-based alignments have been used, with sequences from as few as two species to many dozens. In addition to the SCFG/CM approach outlined above, various groups have based structure inference on pair-models, on folding energy, on the powerful (but computationally expensive) Sankoff dynamic programming approach and heuristic approximations to it, on machine learning approaches, and combinations of these. Some approaches attempt to exploit phylogeny, others do not. Some have attempted clustering or other broader integration of results. For example, Lu et al. [62] combined RNA structure prediction with modENCODE data including RNAseq to characterize ncRNAs in *Caenorhabditis elegans*. Another recent approach makes implicit prediction through an SVM-based classifier tackling the problem of distinguishing structured from nonstructured UTRs [63]. Amidst this great diversity of approaches, one point of agreement is that these methods all identify hundreds to thousands of novel ncRNA candidates. Table 1 summarizes a sample of these results.

Table 1
Some genomic ncRNA scans

METHOD[a]	ORGANISMS[b]	NUM[c]	NOTES[d]
QRNA [71] (pair-SCFG)	Bacteria	275	2-way, Seq, SW, [72]
	Yeasts	92	2-way, Seq, SW, [73]
RNAz [43] (Energy, SVM)	*Ciona*	2,109	2-way, Seq, SW, [74]
	Worm	2,366	2-way, Seq, SW, [75]
	Vertebrates/ PhastCons	35,985	8-way, Seq, SW, [76]
	Vertebrates/ ENCODE	3,707	28-way, Seq, SW, [77]
EvoFold [40] (phylo-SCFG)	Vertebrates/ PhastCons	48,479	8-way, Seq, SW, [40]
	Vertebrates/ ENCODE	4,986	28-way, Seq, SW, [77]
Dynalign [29] (Sankoff, SVM)	Bacteria	995	2-way, Seq, SW, H, [78]
FOLDALIGN [79, 80] (Sankoff)	Human-mouse	1,297	2-way, Str, [61]
CMfinder [1] (SCFG, EM)	Bacteria	1,466	*n*-way, Str, [2]
	Vertebrates/ ENCODE	6,587	17-way, Str, [4]
EvoFam [70] (SCFG)	Vertebrates/ PhastCons	220	41-way, Seq, [70]

[a] Name, references, and core methodology

[b] Genomes screened: ENCODE means pilot ENCODE regions (≈ 1% of the human genome, plus orthologous regions of other vertebrates); PhastCons means PhastCons conserved regions (≈ 5% of the human genome, plus orthologous regions of other vertebrates)

[c] Number of predicted RNA structures; where possible, numbers quoted are from a "stringent" set reported in the original paper, e.g., "$p > 0.9$" for RNAz and Dynalign, "Top 50%" for EvoFold ENCODE scan, and partially hand-curated for EvoFam. It is important to note that follow-up validation of predictions has been limited and estimated false discovery rates are often high

[d] *Notes*: *2-way* pairwise alignments, others were multiple alignments (CMfinder bacterial scan used varying numbers of sequences; RNAz generally selects five sequences if more are available), *Seq* sequence-based alignments, *Str* local structure-based alignments, *SW* limited length, sliding windows, *H* HMM-based alignment constraints

5 Conclusions and Future Prospects

To summarize, key challenges in *de novo* discovery of novel noncoding RNA genes (and other functional ncRNA elements, like riboswitches) center on the problems of (a) finding a sufficiently large set of putative representatives, sufficiently enriched with *actual* representatives, and (b) successfully inferring the consensus features of the family from these candidates, given that the candidate set is almost certainly of suboptimal diversity, with ill-defined borders, and contaminated with non- and atypical examples. "Suboptimal diversity" includes situations where sequence conservation is too high, potentially causing failures due to lack of covariation and/or spurious inclusion of "consensus" features that are absent from the broader family. It also includes situations where sequence conservation is too low, causing poor alignments and/or reduced enrichment of actual examples.

The case studies presented above outline some approaches to answering these challenges. Of these, we single out three general features that seem noteworthy and broadly applicable. First is the importance of exploiting prior knowledge where possible, especially for the purposes of generating candidate sequence sets on which to apply RNA motif discovery algorithms. The selection of upstream sequences of genes containing orthologous conserved domains for riboswitch discovery is one such example. A second important strategy is the integration of discovery with search—exposing additional examples of any motif helps both in accurately characterizing the motif and potentially in the ultimate functional characterization of the RNA. Again, this strategy has proved very useful in riboswitch discovery in prokaryotes, and we expect it to play an increasingly important role in analysis of mammalian genomes, although their larger sizes pose significant computational difficulties. Thirdly, we think that iterative approaches will continue to play important roles in these analyses. The enormous computational cost of "exact" algorithms even for idealized versions of these problems (e.g., the Sankoff algorithm) seem to necessitate use of heuristic approximations, but iterative refinement of the initial solution (e.g., as in CMfinder's EM-like approach, or the integration of search with discovery as mentioned above) builds in some tolerance for errors that will inevitably be made in the early stages, thus allowing them to be fast enough to be feasible.

Many challenges remain. Two issues are paramount among these. First, all genome-scale approaches to date, as assessed by the sophisticated methods mentioned in Subheading 3.4, have been plagued with high false discovery rates and/or low sensitivity. Better scoring, more realistic null models, and deeper exploitation of phylogenetic information might all help, as would more refined motif inference. As one example of the latter, improvements to

CMfinder's initialization might be attainable by augmenting single sequence folding with pairwise structural alignments, inclusion of suboptimal structural alignments, and/or phylogenetic information. The second major issue is that computational cost remains very high for the algorithms in use today. For example, one of the smaller projects mentioned above, the CMfinder Firmicutes scan, used about 1 year of computer time, and genome-scale screens in vertebrates have used hundreds. Obviously, parallelization is possible, computer costs continue to decline, and, independently, algorithms get faster. Dramatic improvements in covariance model search speed, for example, have been obtained in the past few years [19, 64–67]. Nevertheless, these problems are still at the margins of affordability. Genome-scale clustering of results [68–70] is another area that is potentially very compute-intensive and has received relatively little attention to date. We hope for progress on all of these fronts.

Acknowledgements

This work is supported by the Danish Council for Independent Research (Technology and Production Sciences), the Danish Council for Strategic Research (Programme Commission on Strategic Growth Technologies), as well as the Danish Center for Scientific Computing.

References

1. Yao Z, Weinberg Z, Ruzzo WL (2006) CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 22(4):445–452. PMID:16357030
2. Yao Z, Barrick J, Weinberg Z, Neph S, Breaker R, Tompa M, Ruzzo WL (2007) A computational pipeline for high-throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. PLoS Comput Biol 3(7):e126. PMID:17616982
3. Weinberg Z, Barrick JE, Yao Z, Roth A, Kim JN, Gore J, Wang JX, Lee ER, Block KF, Sudarsan N, Neph S, Tompa M, Ruzzo WL, Breaker RR (2007) Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. Nucleic Acids Res 35:4809–4819. PMID:17621584
4. Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, Gorodkin J (2008) Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. Genome Res 18:242–251. PMID:18096747
5. Gorodkin J, Knudsen B (2000) RNA informatik. Naturens Verden 11–12:2–9
6. Eddy SR (2001) Non-coding RNA genes and the modern RNA world. Nat Rev Genet 2(12):919–929. PMID:11733745
7. Eddy SR (2002) Computational genomics of noncoding RNA genes. Cell 109(2):137–140. PMID:12007398
8. Bompfünewerer AF, Flamm C, Fried C, Fritzsch G, Hofacker IL, Lehmann J, Missal K, Mosig A, Müller B, Prohaska SJ, Stadler BMR, Stadler PF, Tanzer A, Washietl S, Witwer C (2005) Evolutionary patterns of non-coding RNAs. Theory Biosci 123(4):301–369. PMID:18202870
9. Mattick JS, Makunin IV (2006) Non-coding RNA. Hum Mol Genet 15(1):R17–R29. PMID:16651366
10. Bompfünewerer AF, Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, Prohaska SJ, Rose D, Stadler PF, Tanzer A, Washietl S, Will S (2007) RNAs everywhere: genome-wide annotation of structured RNAs.

J Exp Zoolog B Mol Dev Evol 308:1–25. PMID:17171697
11. Gorodkin J, Hofacker IL, Torarinsson E, Yao Z, Havgaard JH, Ruzzo WL (2010) *De novo* prediction of structured RNAs from genomic sequences. Trends Biotechnol 28:9–19 (Feature Review). PMID:19942311
12. Gorodkin J, Hofacker IL (2011) From structure prediction to genomic screens for novel non-coding RNAs. PLoS Comput Biol 7(8):e1002100. PMID:21829340
13. Washietl S, Will S, Hendrix DA, Goff LA, Rinn JL, Berger B, Kellis M (2012) Computational analysis of noncoding RNAs. Wiley Interdiscip Rev RNA 3(6):759–778. PMID:22991327
14. Pace NR, Thomas BR, Woese CR (1999) Probing RNA structure, function, and history by comparative analysis. In: Gesteland RF, Cech TR, Atkins JF (eds) The RNA world, Chap. 4. Cold Spring Harbor Laboratory, Cold Spring Harbor, pp 113–141
15. Shang L, Xu W, Ozer S, Gutell RR (2012) Structural constraints identified with covariation analysis in ribosomal RNA. PLoS One 7(6):e39383. PMID:22724009
16. Barrick JE, Breaker RR (2007) The distributions, mechanisms, and structures of metabolite-binding riboswitches. Genome Biol 8(11):R239. PMID:17997835
17. Zuker M (1989) Computer prediction of RNA structure. Methods Enzymol 180:262–288. PMID:2482418
18. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. Monatshefte für Chemie 125:167–188
19. Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335–1337. PMID:19307242
20. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge
21. Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? Nat Biotechnol 26(8):897–899. PMID:18688245
22. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs in MEME. In: Proceedings of the third international conference on intelligent systems for molecular biology. AAAI, Menlo Park, pp 21–29. PMID:7584439
23. Eddy SR, Durbin R (1994) RNA sequence analysis using covariance models. Nucleic Acids Res 22(11):2079–2088. PMID:8029015
24. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, Underwood RC, Haussler D (1994) Stochastic context-free grammars for tRNA modeling. Nucleic Acids Res 22(23):5112–5120. PMID:7800507
25. Touzet H, Perriquet O (2004) CARNAC: folding families of related RNAs. Nucleic Acids Res 32(Web server issue):W142–W145. PMID:15215367
26. Ji Y, Xu X, Stormo GD (2004) A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. Bioinformatics 20(10):1591–1602. PMID:14962926
27. Sankoff D (1985) Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math 45:810–825
28. Gorodkin J, Heyer LJ, Stormo GD (1997) Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res 25(18):3724–3732. PMID:9278497
29. Mathews DH, Turner DH (2002) Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. J Mol Biol 317(2):191–203. PMID:11902836
30. McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 29:1105–1119. PMID:1695107
31. Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. J Mol Biol 319(5):1059–1066. PMID:12079347
32. Altschul SF, Erickson BW (1985) Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. Mol Biol Evol 2(6):526–538. PMID:3870875
33. Babak T, Blencowe BJ, Hughes TR (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. BMC Bioinformatics 8:33. PMID:17263882
34. Gesell T, Washietl S (2008) Dinucleotide controlled null models for comparative RNA gene prediction. BMC Bioinformatics 9:248. PMID:18505553
35. Anandam P, Torarinsson E, Ruzzo WL (2009) Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. Bioinformatics 25:668–669. PMID:19136551
36. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22(21):2688–2690. PMID:16928733
37. Gowri-Shankar V, Rattray M (2007) A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. Mol Biol Evol 24(6):1286–1299. PMID:17347157

38. Knudsen B, Hein J (1999) RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. Bioinformatics 15(6):446–454. PMID:10383470
39. Knudsen B, Hein J (2003) Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res 31(13):3423–3428. PMID:12824339
40. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, Haussler D (2006) Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2:e33. PMID:16628248
41. Yao Z (2008) Genome scale search of non-coding RNAs: bacteria to vertebrates. Ph.D. thesis, Department of Computer Science and Engineering, University of Washington
42. Bernhart SHF, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinformatics 9:474. PMID:19014431
43. Washietl S, Hofacker IL, Stadler PF (2005) Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 102:2454–2459. PMID:15665081
44. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR (2003) Rfam: an RNA family database. Nucleic Acids Res 31(1):439–441. PMID:12520045
45. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33(Database issue):121–124. PMID:15608160
46. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A (2009) Rfam: updates to the RNA families database. Nucleic Acids Res 37(Database issue):D136–D140. PMID:18953034
47. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, Bateman A (2011) Rfam: Wikipedia, clans and the "decimal" release. Nucleic Acids Res 39(Database issue):D141–D145. PMID:21062808
48. Marchler-Bauer A, Anderson JB, Cherukuri PF, DeWeese-Scott C, Geer LY, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2005) CDD: a Conserved Domain Database for protein classification. Nucleic Acids Res 33(Database issue):D192–D196. PMID:15608175
49. Weinberg Z, Regulski EE, Hammond MC, Barrick JE, Yao Z, Ruzzo WL, Breaker RR (2008) The aptamer core of SAM-IV riboswitches mimics the ligand-binding site of SAM-I riboswitches. RNA 14:822–828. PMID:18369181
50. Regulski EE, Moy RH, Weinberg Z, Barrick JE, Yao Z, Ruzzo WL, Breaker RR (2008) A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. Mol Microbiol 68:918–932. PMID:18363797
51. Sudarsan N, Lee ER, Weinberg Z, Moy RH, Kim JN, Link KH, Breaker RR (2008) Riboswitches in eubacteria sense the second messenger cyclic di-GMP. Science 321(5887):411–413. PMID:18635805
52. Wang JX, Lee ER, Morales DR, Lim J, Breaker RR (2008) Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. Mol Cell 29:691–702. PMID:18374645
53. Meyer MM, Roth A, Chervin SM, Garcia GA, Breaker RR (2008) Confirmation of a second natural preQ1 aptamer class in Streptococcaceae bacteria. RNA 14:685–695. PMID:18305186
54. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, Breaker RR (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome Biol 11(3):R31. PMID:20230605
55. Weinberg Z, Perreault J, Meyer MM, Breaker RR (2009) Exceptional structured noncoding RNAs revealed by bacterial metagenome analysis. Nature 462(7273):656–659. PMID:19956260
56. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W (2004) Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 14(4):708–715. PMID:15060014
57. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome Res 12(6):996–1006. PMID:12045153
58. ENCODE Project Consortium et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816. PMID:17571346
59. Lunter G, Ponting CP, Hein J (2006) Genome-wide identification of human functional DNA using a neutral indel model. PLoS Comput Biol 2(1):e5. PMID:16410828
60. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment

programs upon structural RNAs. Nucleic Acids Res 33(8):2433–2439. PMID:15860779
61. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J (2006) Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. Genome Res 16(7):885–889. Erratum: Genome Res 16:1439, 2006. PMID:16751343
62. Lu ZJ, Yip KY, Wang G, Shou C, Hillier LW, Khurana E, Agarwal A, Auerbach R, Rozowsky J, Cheng C, Kato M, Miller DM, Slack F, Snyder M, Waterston RH, Reinke V, Gerstein MB (2011) Prediction and characterization of noncoding RNAs in *C. elegans* by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res 21(2):276–285. PMID:21177971
63. Chen XS, Brown CM (2012) Computational identification of new structured cis-regulatory elements in the 3′-untranslated region of human protein coding genes. Nucleic Acids Res 40(18):8862–8873. doi: 10.1093/nar/gks684. PMID:22821558
64. Weinberg Z, Ruzzo WL (2004) Faster genome annotation of non-coding RNA families without loss of accuracy. In: RECOMB04: Proceedings of the eighth annual international conference on computational molecular biology. ACM, San Diego, pp 243–251. http://doi.acm.org/10.1145/974614.974647
65. Weinberg Z, Ruzzo WL (2004) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. Bioinformatics 20(1):i334–i341. PMID:15262817
66. Weinberg Z, Ruzzo WL (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. Bioinformatics 22(1):35–39. PMID:16267089
67. Sun Y, Buhler J, Yuan C (2012) Designing filters for fast-known ncRNA identification. IEEE/ACM Trans Comput Biol Bioinformatics 9(3):774–787. PMID:22084145
68. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol 3(4):e65. PMID:17432929
69. Tseng HH, Weinberg Z, Gore J, Breaker RR, Ruzzo WL (2009) Finding non-coding RNAs through genome-scale clustering. J Bioinformatics Comput Biol 7:373–388. PMID:19340921
70. Parker BJ, Moltke I, Roth A, Washietl S, Wen J, Kellis M, Breaker R, Pedersen JS (2011) New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. Genome Res 21(11):1929–1943. PMID:21994249
71. Rivas E, Eddy SR (2001) Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics 2(1):8. ISSN 1471-2105. PMID:11801179
72. Rivas E, Klein RJ, Jones TA, Eddy SR (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 11(17):1369–1373. PMID:11553332
73. McCutcheon JP, Eddy SR (2003) Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. Nucleic Acids Res 31(14):4119–4128. PMID:12853629
74. Missal K, Rose D, Stadler PF (2005) Non-coding RNAs in *Ciona intestinalis*. Bioinformatics 21(2):ii77–ii78. PMID:16204130
75. Missal K, Zhu X, Rose D, Deng W, Skogerbø G, Chen R, Stadler PF (2006) Prediction of structured non-coding RNAs in the genomes of the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. J Exp Zoolog B Mol Dev Evol 306(4):379–392. PMID:16425273
76. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. Nat Biotechnol 23(11):1383–1390. PMID:16273071
77. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, Lindemeyer M, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigo R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF (2007) Structured RNAs in the ENCODE selected regions of the human genome. Genome Res 17(6):852–864. PMID:17568003
78. Uzilov AV, Keegan JM, Mathews DH (2006) Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 7:173. PMID:16566836
79. Havgaard JH, Lyngsø RB, Stormo GD, Gorodkin J (2005) Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. Bioinformatics 21(9):1815–1824. PMID:15657094
80. Havgaard JH, Torarinsson E, Gorodkin J (2007) Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol 3:1996–1908. PMID:17937495